I)Title:

Browser Event-Stream Data Mining
for Web Content Syndication

II) Introduction:

Controlled and accurate rendering of web resources has grown into a necessity nowadays, to support mass of surfers in browsing the Internet. This is due to the vastness of Internet and scatter of information in web pages over the Internet. Browsing the internet has become a very tedius job, especially for a layman. Also it consists of added overhead of 'uncertainty in reaching the desired end' on the surfer. Thus, automization and personalization of web usage is a very popular area of research today.

Personalization includes the feature of automatically adapting the browser so that it can learn the next information expected by a surfer and render the relevant web pages without obtaining an explicit query from the web user. This feature is desirable by any website and search engine which are being heavily used by web users. Personalization may be achieved by –

a) Understanding the next information need by a web user.
b) Automatically generating the required web pages.

An understanding of the webpage requirement by a surfer may be achieved by learning their browsing pattern or behavior over the world wide web. A website (a collection of numerous webpages) or multiple websites are being continuously visited by huge number of people, hence Data Mining (Web Mining) is the obvious solution for learning the visibly unpredictable behavior of web users. This branch of web mining is called Web Usage Mining. This research work aims at web usage mining in order to find a way to achieve personalization of web usage.

This research work is a part of an Indo-German project in Web Optimization. The data of surfers' behavior is being available from our German counterpart. Hence, in this work aims Web Usage Mining is being studied on the above data to learn the browsing behavior of surfers.

Data mining is a technique of churning out significant patterns from huge amount of data. Data mining applied on web based data is called Web Mining. The basic techniques in data mining are: Association Rule Mining(ARM), Classification, and Clustering. ARM is a technique of discovering rules that pass certain constraints in a dataset. Classification technique is about constructing a model from the classified objects to classify future objects. This process involves prediction of future class labels, whereas ARM involves only the description of relationship among items in a database. ARM approach 'Apriori' and classification approach 'C4.5' has been studied for our purpose. Associative classification is a new approach that integrates ARM and classification technique. Some of them studied are CBA[1,2], CMAR[3], CPAR[4], MMAC[5]. They have shown almost similar accuracy. Studies about their suitability on the web data is our ongoing work for achieving our goal. There are factors like size of dataset, attribute types, number of attributes, etc. which affect the accuracy and performance of any data mining approach. Clustering is a technique that groups various elements into groups without prior knowledge of categories existing in the dataset.

III) A Brief Review Of Work Already Done In The Field:

In Web Usage Mining, study is about behaviour of surfer. Previous works show that

I. Web Personalization has been approached in three ways –

Approaches of Web Personalization

Profiling of surfers to understand his preferences, so that web pages can be prefetched for him. [18,23,25]	**Popular paths of web pages** that are frequently visited were traced to predict the next expected web page, so that highly probable web pages could be prefetched. [16,17,19,24]	**Understanding the browsing behaviour** of surfer so that type of data being aimed can be predicted and related web locations can be fetched. [13,14,15]

II. Data Mining Techniques that has been explored till now show that –

In the *first* approach, Profiling used the method of creating rule based profiles, content based profiles and collaborative filtering. But data mining was not used in this approach.

In the *second* approach, Clustering was on 'path of visited web pages' based on similarity of the paths, along with weights on a web page based on time spent it. Improvements on this method was also done using frequency of cache reference, link access, longest common subsequence and Association Rule Mining.

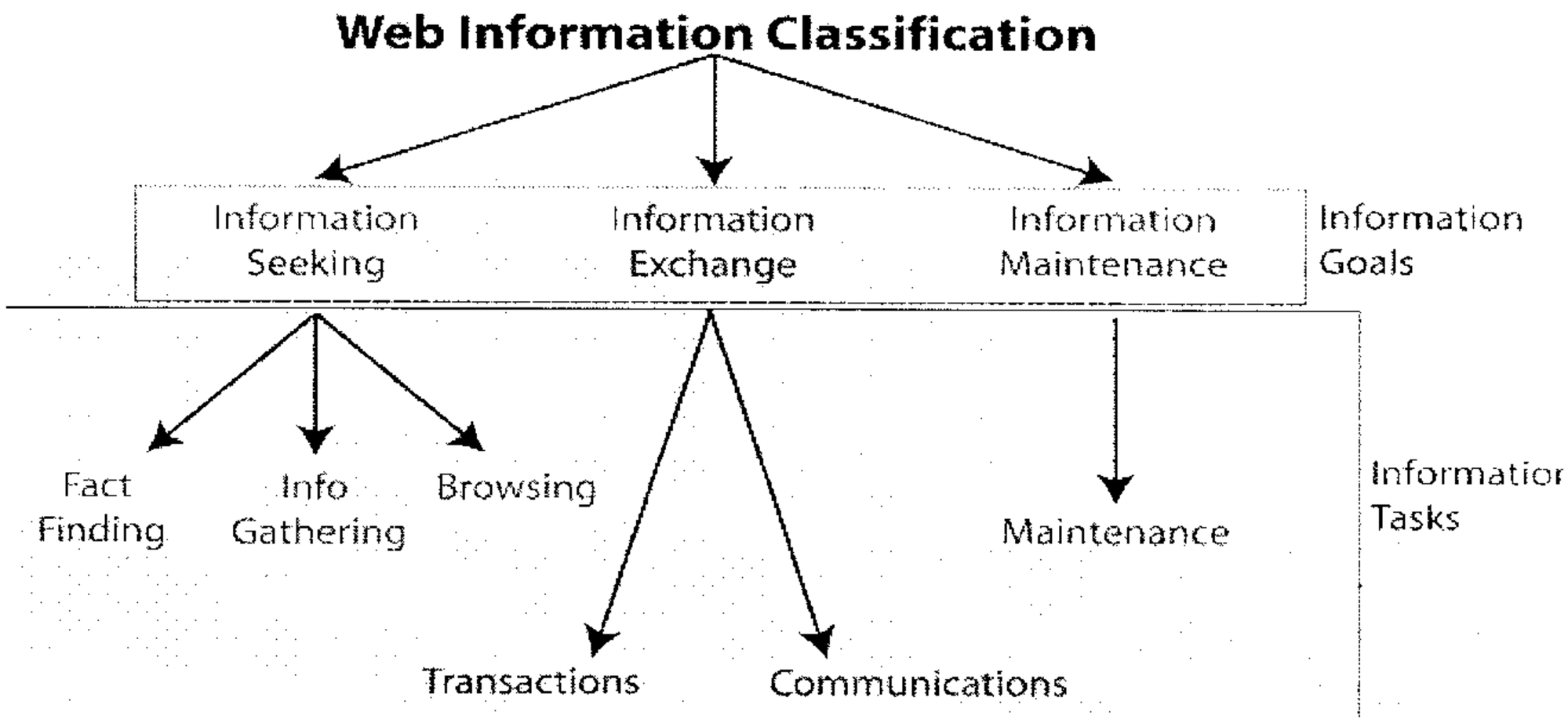
In the *third* approach, some parameters of browsing behaviour was recorded in the laboratory using customized browsers. Expert opinions and pilot groups were used to finalize categories of behaviour. Statistical analysis was done to find the characteristics of the categories of the behaviour. Till now, no data mining approach is found in this line of research.

M. Kellar et.al.[14] has made a detailed study and provided the following tasks along with their characteristics at the browser-

GENERAL TASKS

Tasks → ***Characteristics ↓***	**FACT FINDING**	**INFORMATION GATHERING**	**BROWSING**	**TRANSAC _TIONS**
Duration	Short	Long	Short	Short
No. of pages viewed	Small	Large	Small	(depends)
Search component	Large	Large	Often repeated	Most often repeated
Queries	Long	Short		
Usage of browser functions	Little	Greatest	little	Little
Typed in URLs	Y	Y	Y	Y
Google Toolbar	Y	Y	Y	Y
Others	Bookmarks	Auto-complete	Select-URL	

The authors extended their work [15] at understanding of the high level tasks of the surfers. They developed a classification of web information tasks as follows –



Anne Gutschmidth et.al., our German collaborator made a study [13] to formulate hypotheses describing task-dependent behavioral patterns and mainly concentrated on three user tasks – Fact Finding, Just Browsing and Information Gathering. They found differences between the three user tasks as follows –

- Fact Finding shows a smaller average page view duration than Just Browsing and Information Gathering.
- Users performing Fact Finding look at more pages during one minute than with Information Gathering and Just Browsing.
- Fact Finders do not tend to page revisits.
- Users performing Information Gathering spend little time on the start page of the newspaper compared to Fact Finding and Just Browsing.
- O.P.Vyas et.al.[12] has proposed the Temporal Associative Classification for useful data mining tasks on transactional database. It is being proposed to make use of this approach for click-stream data mining.

IV) Noteworthy Contributions In The Field Of Proposed Work:

The approach of understanding the behavior of surfer at browsing desk, is relatively a new emerging field and have few laboratory contributions in this direction. Contribution of data mining have not been found so far.

In [WUM], we have explored Association rule Mining algorithm Apriori and Classification Rule Mining algorithm C4.5, which is a standard benchmark algorithm and received encouraging results. Apriori was used in order to find if any related group of task measures left out as unobserved. Large number of patterns were produced. The frequent items indicated of many relations showing that users showing one kind of behavior at browser will also show another kind of behavior.

C4.5 was applied on the web data and significant classification rules with 10.0% errors both before pruning and after pruning of the tree was obtained. The results are encouraging to use improved dataset and continue this approach for further study in creating a model for prediction of web user's task.

In [DAC], the hybrid data mining technique Associative Classification Rule Mining which is more accurate and efficient in terms of memory and speed, is being explored and expected to give good results. In this paper, I have made a study of the data structures used in various most accurate Associative classifiers such as CBA, CMAR, CPAR and MMAC to explore their suitability on different types of dataset. It is found that CBA involves much resource and time-consuming steps as it has multiple passes of Apriori ARM, which is itself a multi-pass approach. CMAR uses FP-growth approach of ARM and then creates a CR-tree of rules of similar structure. Hence it is very efficient both in storage and rule retrieval. It is useful for large databases. CPAR also requires O(nk) time for every rule generation and requires lot of space if the no of items are large. It generates PN-array, whose size is proportional to the no of items in the database. Similarly, MMAC also creates a list of size equal to the number of items in the database, but the training data requires only one scanning. So it is very efficient in terms of time and storage.

Currently, for the above purpose, results of the above associative classifiers is being explored on the above data and results are being analyzed to derive inferences about the browsing tasks.

V) Proposed Methodology During The Tenure Of Research Work:

In our endevour to understand the behavior of user while browsing the internet, pattern using the browser requires a series of task. As we have division of work with our German collaborator. They are working in laboratory to record the pattern of behavior of various users.. This work is aimed to explore data mining approaches on the above recorded data. We are to work in the following ways –

1. Pre_processing of Data: The collected data may contain many errors or missing values. Hence, the data need to be treated for missing, noisy and outlier values in it, so that results obtained on application of data mining may be correct and consistent.

2. Usage of Associative Classification approach[12] on the above data in order to obtain the temporal characteristics of known categories of tasks at the browser.

3. Derive inferences about the characteristics of the browsing tasks and explore the possibly unknown relationship existing among the tasks' behaviours.

4. Explore more data mining techniques on the browser usage data to find unknown tasks or behavioral aspects existing based on the browsing mechanisms of the surfer.

5. Analysis of the above result for its performance, accuracy and relevance in achieving our goal. Factors like size of dataset, attribute types, number of attributes, etc. will be taken into consideration for analysis of performance existing data mining algorithms.

6. Existing data mining algorithms may be adapted to obtain a complete result about browsing behaviours.

7. New algorithm may be developed that could optimize the results and achieve the above goal.

8. Our proposed algorithm will explore the digital data repository of a e-newspaper being updated on regular basis.

VI) Expected Outcome Of Proposed Work:

The proposed research work is expected to yield some significant finding s and insight in the web content syndication by using the appropriate data mining techniques on the clickstream data. The expected outcome of the proposed work may be as follows –

- Appropriate combination and pattern of browsing mechaniams is proposed to be found.
- Behavioural differences of surfers due to difference in age, expertise level, physical challenge of surfer, etc is proposed to be found.
- Find behavioural differences due to time and occasions such as festival, disaster, etc.
- Development of a new data mining algorithms.
- Based on our work, a new browsing personalization tool is proposed to be designed or developed.

VII) References

[1] B Liu, W Hsu, Y Ma, "Integrating classification and association rule mining" - Proceedings of the KDD, 1997, New York, NY, pp. 80-86, 1998

[2] D Janssens, G Wets, T Brijs, K Vanhoof, *"Integrating Classification and Association Rules by proposing adaptations to the CBA Algorithm"* - Proceedings of the 10th International Conference , 2003

[3] Wenmin Li, Jiawei Han, Jian Pei, *"CMAR: Accurate and Efficient Classification Based on Multiple Class- Association Rules"* – Proceedings ICDM Nov,2001, pp369-376, San Jose, CA.

[4] Xiaoxin Yin, Jiawei Han – *"CPAR: Classification based on Predictive Association Rules"*. Proceedings of SIAM Int. Conf. on Data Mining (SDM'03), San Fransisco, CA, pp.331-335.

[5] F Thabtah, P Cowling, Y Peng , *"Multiple labels associative classification"* Knowledge and Information Systems, Vol. 9, No. 1. (2006), pp. 109-129.

[6] F Thabtah, P Cowling, Y Peng , *"A Study of Predictive Accuracy for Four Associative Classifiers"* - Journal Of Digital Information Management, 2005

[7] J. R. Quinlan and R. M. Cameron-Jones. *"FOIL:A midterm report."* In Proc. 1993 European Conf. Machine Learning, pp. 3{20, Vienna, Austria, 1993.

[8] F Thabtah. *"Rules Pruning in Associative Classification Mining."* Proceedings of the IBIMA Conference, 2005

[9] F Coenen, P Leng, L Zhang."*Threshold Tuning for Improved Classification Association Rule Mining."*; Lecture Notes in Artificial Intelligence, 2005 – Springer

[10] F Coenen, P Leng. "*Obtaining Best Parameter Values for Accurate Classification*" Data Mining, Fifth IEEE International Conference on, 2005

[11] Yanbo J. Wang, Qin Xin, Frans Coenen. "*A Novel Rule Ordering Approach in Classification Association Rule Mining*". P. Perner (Ed), Proceedings of the fifth International conference on Machine Learning and Data Mining (MLDM-07), Leipzig, Germany, july 2007. LNAI 4571, Springer-verlag Berlin Heidelberg. 2007, p. 339-348

[12] To be presented: O.P.Vyas, Ranjana Vyas, Vivek Ranga, Anne Gutsschmidt. "*Exploring Temporal Associative Classifiers for Business Analytics"* First IIMA conference on Advanced Data Analysis, business Analytics & Intelligence. June 6-7, 2009, Ahmedabad, India.

[13] A. Gutschmidt, C.H. Cap, F.W. Nerdinger. "*Paving the Path to Automatic User Task Identification*" Workshop on Common Sense Knowledge and Goal-Oriented Interfaces on the International Conference on Intelligent User Interfaces 2008.

[14] M Kellar, C Watters, M Shepherd. " *A Goal-based Classification of Web Information Tasks"*; Proceedings of ASIS&T 2006

[15] Melanie Kellar, Michael Shepherd, Carolyne Waters. "*A Field Study Characterizing Web based Information Seeking Tasks.*" A Technical Report. University Ave, Canada 2005.

[16] Sule Gunduz, M.Tmerozsu. "*A Web Page Prediction Model Based on Click stream Tree Representation of User Behaviour*" SIGKDD'03-2003

[17] J.Srivastava, R. Cooley, M. Deshpande and P.N.Tan. " *Web Usage Mining: Discovery and Application of Usage Patterns from Web Data*". ACM SIGKDD explorations, 2000

[18] B Mobasher, R Cooley, J Srivastava. " *Automatic Personalization Based on Web Usage Mining*". Communications of the ACM, 2000.

[19] Robert Cooley, Bamshad Mobasher, Jaideep Srivastava " *Data Preparation for Mining World Wide Web Browsing Patterns*". Knowledge and information Systems, 1(1), 1999

[20] Data Mining-Multimedia, SoftComputing and BioInformatics; by Sushmitra Acharya and Tinku Acharya. A John Wiley & Sons, Inc., Publication; 2003

[21] Data Mining- Introductory and Advanced Topics; by Margaret H. Dunham. Pearson Education; 2003

[22] Data Mining: Concepts and Techniques; by Jiawei Han & Micheline Kamber. Second Edition; Morgan Kaufmann Publishers

[23]Sander Spek. "*Personalization*". *Institute of Knowledge Agent Technology, University of Maastricht(IKAT).*

[24] Petr Berka[3], Vladimír Laš[3] and Tomáš Kočka[3], "*Rule Induction for Click-Stream Analysis: Set Covering and Compositional Approach*". Advances in Soft Computing Book Series 2005

[25] Ibrahim Cingil, Asuman Dogac, Ayca Azgin. "*A Broader Approach to Personalization*". Communication of the ACM. Date: 2000. Volume: 43. Issue: 8

VIII) List Of Published Papers Of The Candidate

[FS1] Kavita Das, "Fuzzy Sets". *Proceedings of National Conference; Forum of Scientists & Technologists (FOST), Sep 24, 2006, Raipur.Chhattisgarh*

[MDM] O.P. Vyas, M.K. Tiwari, Kavita Das, "Role of MANET in Disaster Management".- Extended Abstract. *Proceedings of National Conference; INDIACom-2007 Computing for Nation Development, Feb 23-24 2007, Bharati Vidyapeeth's Institute of Computer Applications & Management, New Delhi.*

[DAC] Kavita Das, O.P. Vyas, " Overview of Data Structures in Associative Classifiers". *Proceedings of National Conference; Colloquium on Mobile Communication and Data & Knowledge engineering-2009, Jan 2-3 2009, Pt. Ravishankar Shukla University, Raipur, Chhattisgarh*

[WUM] Kavita Das, O.P. Vyas, " Suitability of Web Usage Mining for Web Content Syndication". *Proceedings of National Conference; INDIACom-2009 Computing for Nation Development, Feb 26-27 2009, Bharati Vidyapeeth's Institute of Computer Applications & Management, New Delhi.*

Signature of Supervisor

(Prof. O. P. Vyas)

S.O.S IN COMPUTER SCIENCE;
Pt. Ravishankar Shukla University,
Raipur (C.G.)

Signature of Candidate

(Kavita Das)

Forwarded

Head
S.O.S in Computer Science
Pt. Ravishankar Shukla University,
Raipur (C.G.)

HEAD,
S.O.S IN COMPUTER SCIENCE
PT. RAVISHANKAR SHUKLA UNIVERSITY
RAIPUR (C.G.)